先端科学技術研究科　修士論文要旨

所属研究室 (主指導教員)	コンピューティング・アーキテクチャ (中島　康彦 (教授))		
学籍番号	2411038	提出日	令和 8年 1月 16日
学生氏名	衛藤　優		
論文題目	Implementation and Evaluation of Weight-Quantized LLMs on a CGLA Architecture CGLAアーキテクチャにおける重み量子化大規模言語モデルの実装と評価		

要旨

Large Language Models (LLMs), such as LLaMA, are computationally intensive and memory-heavy, raising significant concerns regarding energy consumption. This thesis presents the implementation and evaluation of weight-quantized LLMs on IMAX3, a novel Coarse-Grained Linear Array (CGLA)-based in-memory accelerator architecture designed for both high-performance computing and low-power edge applications. By porting llama.cpp—a framework optimized for quantized inference—to IMAX3, this study enables a direct comparative analysis of performance and energy efficiency against conventional CPUs and GPUs.

Experimental results demonstrate that while IMAX3 exhibits longer execution times for quantized models (specifically Q8_0 and Q3_K) compared to CPUs and GPUs, it achieves competitive Power-Delay Product (PDP) and Energy-Delay Product (EDP) under specific thread configurations. The study also identifies architectural limitations, notably the absence of ARM control cores and the high power consumption of the large local memory. Future work will focus on optimizing energy efficiency by exploring alternative memory technologies, reducing memory capacity, and integrating additional ARM or Intel cores to enhance the scalability of the IMAX3 prototype.